Available online at www.sciencedirect.com
ScienceDirect
Decision Support Systems 45 (2008) 368–383
Decision Support Systems
www.elsevier.com/locate/dss

Learning Bayesian networks from incomplete databases using a novel evolutionary algorithm

Man Leung Wong [a,*], Yuan Yuan Guo [b,1]

[a] Department of Computing and Decision Sciences, Lingnan University, Tuen Mun, Hong Kong
[b] Department of Computing and Decision Sciences, Lingnan University, Tuen Mun, Hong Kong

Received 10 January 2007; received in revised form 22 January 2008; accepted 24 January 2008
Available online 1 February 2008

Abstract

This paper proposes a novel method for learning Bayesian networks from incomplete databases in the presence of missing values, which combines an evolutionary algorithm with the traditional *Expectation Maximization* (EM) algorithm. A data completing procedure is presented for learning and evaluating the candidate networks. Moreover, a strategy is introduced to obtain better initial networks to facilitate the method. The new method can also overcome the problem of getting stuck in sub-optimal solutions which occurs in most existing learning algorithms. The experimental results on the databases generated from several benchmark networks illustrate that the new method has better performance than some state-of-the-art algorithms. We also apply the method to a data mining problem and compare the performance of the discovered Bayesian networks with the models generated by other learning algorithms. The results demonstrate that our method outperforms other algorithms.

Keywords: Data mining; Machine learning; Bayesian networks; Evolutionary algorithms

1. Introduction

Bayesian networks are popular within the community of artificial intelligence due to their ability to support probabilistic reasoning from data with uncertainty. With a network at hand, probabilistic inference can be conducted to predict the values of some variables based on the observed values of other variables. Hence, Bayesian networks are widely used in many areas, such as decision support systems [33,1], diagnostic and classification systems [27,26,2], information retrieval [21], troubleshooting, data mining [52,25], and so on. Currently, people focus on two kinds of Bayesian network learning problems: parameter learning and structure learning. For parameter learning, a Bayesian network structure is already known to the researchers and the parameters' values of the Bayesian network structure are estimated or optimized. On the other hand, Bayesian network structures have to be discovered from data in structure learning. Existing structure learning methods can be generally classified into two main categories [8]: the dependency analysis approach [49], in which the results of dependency tests are employed to construct a Bayesian network conforming to the findings; and the score-and-search approach [10,20,30], in which a scoring metric is used to evaluate candidate networks while a search method is employed to find a

* Corresponding author. Tel.: +852 26168093; fax: +852 28922442.
E-mail addresses: mlwong@ln.edu.hk (M.L. Wong), yy2guo@gmail.com (Y.Y. Guo).
[1] Tel.: +852 34117460; fax: +852 28922442.

0167-9236/$ - see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.dss.2008.01.002

network structure with the best score. In the latter approach, the score evaluation procedure is time-consuming. Hence, decomposable scoring metrics, such as *Bayesian Information Criterion* (BIC) and *Minimum Description Length* (MDL), are usually used to facilitate the re-evaluation of the score of a network once its structure is changed [30]. Stochastic search methods such as *Genetic Algorithms* (GAs) [32,31,11,52], *Evolutionary Programming* (EP) [51,12], *Estimation of Distribution Algorithms* (EDAs) [6], *Ant Colony Optimization* (ACO) [13], *Scatter Search* [15], and *Hybrid Evolutionary Algorithm* (HEA) [50] have also been proposed in the score-and-search approach. A Bayesian network is usually encoded as an ordered string (consistent with the direction on the graph), or a connection matrix (where the element in the xth row and yth column $c_{xy}=1$ represents that node $x \rightarrow$ node y exists on the graph). Different genetic operators have been designed and employed to find individuals with higher scores. The evolutionary algorithms demonstrate good performance on learning Bayesian networks from complete databases that do not have missing values.

However, learning Bayesian networks from *incomplete* databases, which contain records with missing values or hidden variables, is a difficult problem in real-world applications. Moreover, the patterns of the missing values also affect the performance of the learning methods. Missing values occur in different situations: *missing at random* or *not ignorable* [36,48]. In the first situation, whether an observation is missing or not is independent of the actual states of the variables. Thus the incomplete database may be a representative sample of the complete database. On the other hand, in the second situation, the observations are missing in some specific states for some variables. Different approaches should be adopted for different situations, which again complicates the problem. In this paper, we assume that the unobserved data are missing at random.

Many researchers have been working on parameter learning and structure learning from incomplete databases. For the former, several algorithms can be used, such as Gibbs sampling [19], *Expectation Maximization* (EM) [14,20], and *Bound-and-Collapse* (BC) method [43–45]. For structure learning from incomplete databases, the main issues are how to define a suitable scoring metric and how to search for Bayesian networks efficiently and effectively. Concerning the score evaluation for structure learning, some researchers proposed calculating the expected values of the statistics to approximate the score of candidate networks. Friedman proposed a Bayesian *Structural Expectation Maximization* (SEM) algorithm which alternates between the parameter optimization process and the model search process [16,17]. The score of a Bayesian network is maximized by means of the maximization of the expected score. Peña et al. used the BC + EM method instead of the EM method in their BS-BC + EM algorithm for clustering [40,41]. However, the search strategies adopted in most existing SEM algorithms may not be effective and may make the algorithms find sub-optimal solutions. Myers et al. employed a genetic algorithm to learn Bayesian networks from incomplete databases [38]. Both network structures and the missing values are encoded and evolved. The incomplete database is completed by specific genetic operators during evolution. Nevertheless, it has the efficiency and convergence problems because of the enlarged search space and the strong randomness of the genetic operators for completing the missing values.

In this paper, we propose a novel method that uses EM to handle incomplete databases with missing values and applies an evolutionary algorithm to search for good Bayesian networks. Instead of using the expected values of statistics as in most existing SEM algorithms, our method applies a data completing procedure to complete the database and thus decomposable scoring metrics can be used to evaluate the networks. A strategy is also introduced to get a better initial network from the incomplete database to facilitate the method. We demonstrate that our method outperforms several state-of-the-art algorithms. We also apply the method to a data mining problem and compare the performance of the Bayesian networks obtained by our method with the models induced by several other learning algorithms.

The rest of this paper is organized as follows. In Section 2, we will present the backgrounds of Bayesian networks, the missing value problem, and some Bayesian network learning algorithms. In Section 3, our new method for incomplete databases, *Evolutionary Bayesian Network learning method* (EBN), will be described in details. A number of experiments have been conducted to compare our method with other learning algorithms and the results will be discussed in Section 4. In Section 5, we use our method to discover Bayesian networks from a real-life direct marketing database. We will conclude the paper in the last section.

2. Background

2.1. Bayesian networks

A Bayesian network, G, has a directed acyclic graph (DAG) structure. Each node in the graph corresponds to a discrete random variable in the domain. An edge, $Y \rightarrow X$, on the graph, describes a parent and child relation in